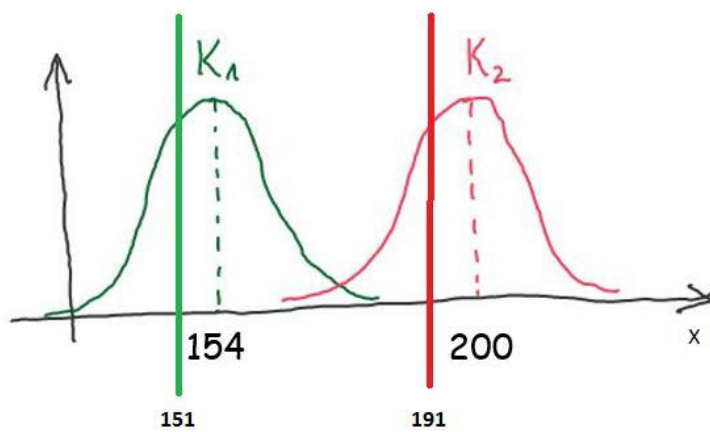


Oblast:

**NAIVNI BAJES KLASIFIKATOR**

**Osnovna ideja:** Izračunati verovatnoće da podatak pripada jednoj ili drugoj klasi. Na osnovu dobijenih vrednosti verovatnoća izvršiti klasifikovanje podatka u neku od klasa. Osnovna pretpostavka jeste da su svi događaji **međusobno nezavisni**.

x	y	z	Region
153	141	125	K1
151	139	123	?
152	140	124	?
153	140	123	?
154	142	126	K1
154	141	124	K1
156	143	126	K1
155	142	125	?
151	138	121	?
155	143	127	?
152	139	122	?
150	138	122	?
197	142	23	?
158	145	128	?
201	146	27	K2
199	144	25	K2
149	136	119	?
156	144	128	?
157	144	127	?



$$P(K1|151) > P(K2|151)$$

$$P(K1|197) < P(K2|197)$$

**Bejsova teorema:** Bajesova teorema se izvodi se direktno iz definicije uslovnih verovatnoća. Uslovna verovatnoća definisana je formulom:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Leva strana formule uslovne verovatnoće čita se na sledeći način: "Verovatnoća da se desi događaj A, ako se prethodno desio događaj B". Obično se neposredno pre dođaja A može desiti više drugih događaja koji utiču na verovatnoću dešavanja dogaja A, pa se ti prethodni događaji obično nazivaju hipotezama i označavaju se sa  $H_i$ . Ali šta ako mi želimo da znamo verovatnoću dešavanja hipoteze ako već znamo da se događaj A dogoditi? Tu se pozivamo na Bajesovu teoremu:

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)}$$

gde je  $i=1,2,\dots,n$ , a  $P(A) > 0$ .

**Primer:** U dve kutije se nalaze kuglice, gde je u prvoj kutiji 25%, a u drugoj kutiji 40% oštećenih kuglica. Na slučajan način se bira kutija i iz nje jedna kuglica. Kolika bi bila verovatnoća da je odabrana neoštećena kuglica?

$H_1$  – Izabrana je prva kutija

$H_2$  – Izabrana je druga kutija

$A$  – Izabrana je neoštećena kuglica

Rešenje ovog problema bi se dobilo formulom uslovne verovatnoće. Ali šta ako je potrebno proveriti **kolika je verovatnoća da je neoštećena kuglica izvučena baš iz prve kutije, ako znamo da je pri jednom slučajnom izvlačenju dobijena neoštećena kuglica?** Rešenje ovog problema se dobija Bajesovom teoremom:  $P(H_1 | A)$ .



## NAIVNI BAJES KAO KLASIFIKATOR SENTIMENTA TEKSTA

Naivna Bajes metoda se jako često koristi za analizu sentimenta tekstualnih sadržaja. Analiza sentimenta je jako složen problem i za ovu namenu se koristi mnogo različitih metoda. Naivni Bajes se pokazuje kao jako dobar, iako je sama metoda jako jednostavna. Pretpostavka naivnog Bajesa jeste da su svi elementi u skupu podataka **nezavisni**. Ako uzmemo tekst i analiziramo reči, očigledno je da one nisu nezavisne ako analiziramo sentiment (npr. „film nije odličan“). Ukoliko reči gledamo nezavisno dobićemo „film“ i „odličan“ u jednoj rečenici što bi moglo da poveća verovatnoću pozitivnog sentimenta. Eksperimentalno je pokazano da pretpostavka o nezavisnosti u analizi teksta ne smanjuje tačnost u velikoj meri, a metoda je zbog pretpostavke o nezavisnosti jako brza.

Prilikom procesiranja sadržaja pisanih prirodnim jezikom, često se koristi tzv. **bag-of-words** model. U tom slučaju se tekstualni sadržaj predstavlja kao **džak** u kome imamo skup reči i broj njihovog pojavljivanja u tekstu. Ceo tekst je iscepan na reči i one su obačene u *džak*. Prilikom korišćenja **bag-of-words** modela možemo primetiti da se ne vodi računa ni o redosledu reči ni o gramatici jezika.

Verovatnoća pripadnosti reči nekoj konkretnoj klasi se računa kao:

$$P(x_i|c) = \frac{N_{ci}}{N_c} \quad (1)$$

Frekvencija pojavljivanja reči se beleži u mapama (Dictionary) u fazi obučavanja klasifikatora, gde je  $N_{ci}$  broj reči  $x_i$  u dokumentima klase  $c$ , a  $N_c$  predstavlja ukupan broj svih reči u dokumentima klase  $c$ . Ova formula ima problem u slučajevima kada se neka reč nikada ne pojavljuje u određenoj klasi. Kako bi se izbegle verovatnoće jedanke nuli, uvodi se takozvano poravnanje. Jedno od najjednostavnijih načina poravnanja je Laplasovo (*Laplacian smoothing*) po kom formula (1) postaje:

$$P(x_i|c) = \frac{N_{ci} + 1}{N_c + |V|} \quad (2)$$

Gde je  $|V|$  ukupan broj reči u rečniku. Na ovaj način se tehnički dodaju virtuelne pojave svih reči, postojale one u rečniku ili ne.

Po Bajesovoj teoremi, verovatnoća da dokument pripada klasi  $c_i$  je data sa:

$$P(c_j|d) = \frac{P(d|c_j) * P(c_j)}{P(d)}$$

Pošto znamo da se dokument sastoji od niza reči i ako usvojimo pretpostavku o nezavisnosti, reči će biti nezavisne jedne od drugih. U tom slučaju dokument možemo predstaviti kao niz reči koje se istovremeno pojavljuju u dokumentu i prethodna formula će umesto dokumenta koristiti niz reči koje se u njemu nalaze:

$$P(c_j|d) = \frac{(\prod P(x_j|c_j)) * P(c_j)}{P(d)} \quad (3)$$

Gde je  $P(c_j)$  verovatnoća da je dokument koji pokušavamo da klasifikujemo baš klase  $c_j$  (tog sentimenta), a  $P(d)$  je verovatnoća da se baš ta reč pojavi u dokumentu čije reči klasifikujemo gledajući sve reči koje klasifikator poznaje. Pošto je cilj klasifikacije da pronade najveću vrednost  $P(c_j|d)$  za sve klase, i  $P(d)$  predstavlja konstantu u tom slučaju, tada je važan samo brojilac u razlomku iz formule 3. Odnosno:

$$P(c_j|d) \propto \left( \prod P(x_j|c_j) \right) * P(c_j)$$

Zbog množenja većeg broja vrednosti manjih od 1, dolazimo do problema premalih razlomljenih brojeva (floating point underflow). Zbog toga se ceo prethodni izraz logaritmuje i dobijamo:

$$P(c_j|d) \propto \left( \sum \ln \left( P(x_j|c_j) \right) \right) + \ln \left( P(c_j) \right) \quad (4)$$

### Zadaci:

1. Otvoriti projekat **NaiveBayes.sln**.
2. *TODO 1:* U direktorijumu **data** se nalazi datoteka **train.tsv**. Potrebno je učitati dati fajl, isparsirati ga i popuniti objekat klase **DataModel**.
3. *TODO 2:* Dopuniti metodu **RemovePunctuation(...)** u klasi **TextUtil** koja iz teksta izbacuje sve znakove interpunkcije. Tekst koji se klasifikuje treba biti pretvoren u bag-of-words model koji ne vodi računa o redosledu reči niti o gramatici. U tom slučaju su znakovi interpunkcije suvišni i potrebno ih je izbaciti.
4. *TODO 3:* Dopuniti metodu **Tokenize(...)** u klasi **TextUtil** koja tokenizuje ulazni tekst, odnosno razdvaja ga na reči (tokene).
5. *TODO 4:* Dopuniti metodu **CountWords(...)** u klasi **TextUtil**, koja broji pojavljivanja svake reči u tekstu. Metoda vraća Dictionary u kome je ključ reč, a vrednost pod tim ključem broj pojavljivanja te reči u tekstu. Ova metoda je poslednji korak u formiranju bag-of-words sa kojim će raditi Naivni Bajes klasifikator.
6. *TODO 5:* Formirati globalni rečnik svih reči koje su se pojavile u svim tekstovima nad kojima se klasifikator obučavao – objekat **vocabulary**. Takođe, formirati rečnike svih reči koje su se pojavljivale u tekstovima koji su bili označeni određenim sentimentom – objekat **word\_count**. *Napomena:* Rečniku svih reči određenog sentimenta se pristupa sa `word_count[sentiment]`.
7. *TODO 6:* Izračunati verovatnoću da je određena reč pozitivnog sentimenta, kao i verovatnoću da je određena reč negativnog sentimenta. Ove verovatnoće se računaju kao odnos broja reči pozitivnog/negativnog sentimenta i ukupnog broja reči koje su klasifikovane. Ove verovatnoće se koriste u narednom koraku klasifikatora -  $P(c_j)$ .
8. *TODO 7:* Izvršiti računanje verovatnoća pripradnosti teksta svakoj od klasa (sentimenata) po formuli datoj u (3).
  - a. 7.1 Izračunati sumu svih logaritama, gde se pod svakom logaritamskom funkcijom nalazi poravnana verovatnoća da se reč pojavljuje u određenoj klasi (verovatnoća se računa po formuli (2)) – prvi deo u formuli (4).
  - b. 7.2 Na rezultat prethodnog računanja dodati logaritam verovatnoće  $P(c_j)$ . Nakon toga se može izračunati verovatnoća sentimenta teksta  $P(c_i|d)$  kao – drugi deo u formuli (4).
9. *TODO 8:* Na konzolu ispisati vrednosti predikcija za pozitivan i negativan sentiment prosledjenog teksta kao i pretpostavku o sentimentu u rečenici.